様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　３　日現在

機関番号：　３２６４２
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2015
課題番号：　２５３３０４１７

研究課題名（和文）潜在意味解析を利用した英語学習語彙リストの生成と評価

研究課題名（英文）Generating Multi-word Expressions for Learners of English as Second Language Using Latent Semantic Indexing

研究代表者

来住　伸子（Kishi, Nobuko）

津田塾大学・学芸学部・教授

研究者番号：５０２４５９９０

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：この研究は、英語学習者に目的や興味に合った教材の自動作成の第一歩として、大規模な英語テキストデータから、学習者が指定した英単語の並びと一緒に使われる可能性の高い用例（文やフレーズ）を自動生成するシステムに関する研究である。潜在意味解析と呼ばれるアルゴリズムを使い、語の出現順が異なる表現や、関連表現を含む用例が生成できるシステムが開発できた。しかし、大規模テキストデータの処理に非常に時間がかかる、英語初級レベルの学習者には不向きな用例を生成する、などの問題点が発見され、システムの再設計や、頻度順などの他のアルゴリズムとの統合が必要な状態である。

研究成果の概要（英文）：This project is aimed at creating a system for generating collocations for the learners of English as a foreign language, from a large English text data set. With this system, the user will be able to enter a list of words, and then get a list of sentences or phrases which are used often with the words that they entered. We completed the system by using an algorithm called a latent semantic indexing(LSI), which is well known for finding synonymous or relevant multi-word expressions. However, the system is very slow in processing a large data set such as Wikipedia data and its output often contains several expressions which are hard to understand for beginning learners of English. We plan to work on improving the system and using other algorithms as well as LSI in the next project.

研究分野：　情報学基礎

キーワード：　潜在意味解析　英語学習　学習コンテンツ開発支援　語彙学習　用例生成

１．研究開始当初の背景

英語を母語としない英語学習者、とくに英語の使用目的が明確な、社会人や大学生の英語学習者が英語を効率的に学ぶには、実際に利用する可能性の高い語彙や用例を使った教材が必要である。そのような英語教材を作成する第一歩としてコーパスの利用が始まっており、出現頻度による語彙リスト作成と編集が実際に行われている。しかし、中級レベル以上の学習者や、特定分野の英語を学びたい英語学習者に向いた語彙リストや、その語彙を学うための用例は、十分には作成されていないのが現状である。

そこで、入手しやすくなった英語テキストデータと、ハードウェアおよびソフトウェア計算資源を活用して、個々の学習者に適した語彙リストとその用例データを自動生成するサービスの構築をめざすことにした。

２．研究の目的

潜在意味解析(Latent Semantic Indexing)と呼ばれるアルゴリズムを利用して、英文大規模テキストデータから語彙リストと用例リストを自動生成する。それらのリストを、既存の英語学習語彙リストと比較し、英語教員による評価を行う。

２．１．潜在意味解析

潜在意味解析とは、大規模行列を特異値分解し、何らかの基準で次数を減らして近似行列を作成し、その近似行列で表現される空間で距離計算などを行う解析方法である。次数の減らし方によって、偶然やランダムな要素の影響を減らし、人間が「意味」と感じる要素をよりはっきり表現する可能性があると考えられている。

大規模テキストデータから、語・文行列を生成し、それに対して潜在意味解析を用いる研究は、１９９０年代前半から Landauer らにより行われている[10]。

２．２．大規模テキストデータ

英語語彙コーパスの整備により、書籍、新聞などで実際に使用された英語テキストデータを容易に入手できるようになった。Brown Corpus や British National Corpus がその先駆者である。現在入手しやすいコーパスは、Corpus of Contemporary American English (COCA) で、約 4.5 億語分のデータが提供されている[1]. また、電子図書館や Wikipedia の普及により、コーパスほど英文品質が高くはないが、最新の話題について記述した英文テキストで Web からすぐに入手できるようになった。

一方、日本国内で使用される、日本人対象の英語教科書は、学習者向きに使用語彙数を減らすなどの大幅な書き換えを行っていたり、扱う話題がかなり古いものであったりすることがある。中級レベル以上の学習者や、特定分野の英語や最近の話題に関する英語を学びたい学習者に向いた教材はあまりない。

この状況を解決するには、大規模英語テキストデータを解析し、学習者の興味や目的に近い語彙や用例を自動生成する技術の確立が必要と考える。本研究では、大規模テキストデータ解析技術として、潜在意味解析に着目し、それを実際のテキストデータに適用した。適用した結果、自動生成した語彙や用例について得た評価やコメントについても報告する。

３．研究の方法

３．１　使用システム

テキストデータの解析システムには、エントリーレベルのサーバーDell PowerEdge R420 E5-2420 を使用した。このサーバーに Python の各種ライブラリをインストールした。Numpy などの標準的なデータ解析ライブラリに加え、gensim を使用した。Gensim ライブラリ[8] は、Python 上のコーパス処理用のベクトル空間処理ライブラリである。潜在意味解析のための行列変換や各種の行列空間での距離計算、類似度計算を行うことができる。大規模行列計算にメモリを利用するだけでなく、ファイルへの書き出しや読み込みも行い、この研究に必要とされる大きな疎行列の計算に適している。また、頻度数, TF-IDF, 潜在意味解析 (LSA) のなどの情報検索用のアルゴリズム各種を提供している。

３．２　使用したテキストデータと処理の主な流れ

利用するテキストデータとして、今回は、コンピュータの歴史に関する書籍と、英語 Wikipedia のアーカイブデータを主に利用した[6]。 前述の COCA は、英語の Wikipedia を含み、Project Gutenberg, Web から集めたデータなど、英語 Wikipedia より大きなテキストデータを提供している[7]。しかし、今回の研究では、サーバーの処理能力などの観点から、日本の英語学習者が読む可能性が高いデータに限定して、処理するデータ量を抑えることにした。Wikipedia からダウンロードしたデータは,Wikipedia Extractor というライブラリ[7]を使い、記事の本文に該当するテキストデータを抽出した。Wikepdia の編集履歴データは含めていない。

これらのテキストデータを nltk ライブラリなどを利用し、カンマや空文を元に文単位に分割した。一文を一文書として、語・文書行列を生成した。通常の文書検索では、語・文書行列を生成するが、本研究では、

Landauer らの研究と同様に、語・文行列をテキストデータから生成した。入力した語の並びから、「文書」ではなく「文（用例）」を生成するためである。そのため、通常の文書検索よりも、非常に大きな行列を扱っている。

４．研究成果

４．１　語のカバー率の調査

英語学習者を対象に、学ぶべき語彙リストがいくつか公開されている。これらの学習語彙リストは、コーパスにおける出現頻度などを利用して作成したものだが、自動生成ではなく、英語教員の意見なども取り入れて編集されている。これらの既存の学習語彙リストの一つ、JACET8000[5]と、実際のテキストデータ（コンピュータの歴史に関する教科書）から生成した語彙リストのカバー率を調査した。結果を図１に示す。

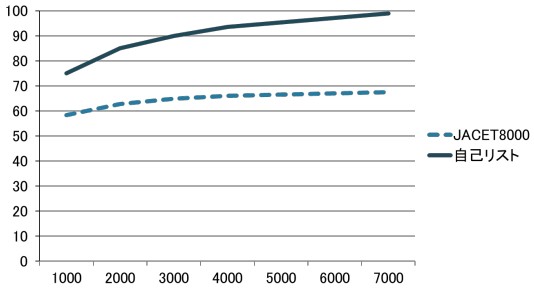

図１　大学教科書における語彙カバー率

図１のグラフの横軸は、語の出現頻度順位を表している。縦軸はカバー率を表している。JACET8000 の最初の１０００語では、６０パーセント程度の語をカバーし、テキスト自身の上位１０００語では、７５パーセント程度の語がカバーされている。JACET8000 の上位７０００語では、このテキストの７０パーセント、テキスト自身の上位７０００語ではほぼ１００パーセントの語彙をカバーしていることが観察できる。つまり、同じ語数、たとえば７０００語学んでも、使用する語彙リストが異なると、知っている語の割合が３０パーセント近く異なることになる。これにより、既製の学習語彙リストより、実際に読んでいるテキストから生成した語彙リストで学んだほうが、学習効率が高いことが予想できる。

４．２　異なるアルゴリズムによって自動生成した用例の比較

次の３種のアルゴリズムで利用して計算した類似度（文と文の距離）を使って、入力した語リストに対して、用例リストを生成した。

- BOW:　語の頻度（bag of words）
- TF-IDF: log(語の頻度)/文が現れる回数
- LSI: 潜在意味解析（ Latent Semantic Indexing).

例えば、語リストとして「user interface」を入力すると、同じテキストデータ（コンピュータの歴史の教科書）から、次のような用例リストが生成された。

BOW によって生成された用例

1. as an operating system dos adequately controlled the machine(‘s) facilities but few would call the user interface intuitive
2. jobs visited the xerox palo alto research center parc in 1979 and came away with whole different idea for user interface
3. windows 3.0 was not really new operating system just user interface program that ran on top of dos

TF-IDF によって生成された用例

1. the program that manages the computer’s resources for the use of other programs launches other computer programs and usually manages the interface with the user
2. windows 3.0 was not really new operating system just user interface program that ran on top of dos
3. as an operating system dos adequately controlled the machine’s facilities but few would call the user interface intuitive

LSI によって生成された用例

1. the program that manages the computer(‘s) resources for the use of other programs launches other computer programs and usually manages the interface with the user
2. windows 3.0 was not really new operating system just user interface program that ran on top of dos
3. parc created computer they called the alto that used bitmapped graphical display graphical user interface gui mouse and programs based on the “what you see is what you get” wysiwyg principle

LSI の用例リストの３番目にでてくる用例が LSI の特徴を表している。３番目の用例は語数が多く、user interface が文の中でもつ重みが BOW や TF-IDF の計算方法では低くなる。しかし、LSI の３番目の用例には、gui(graphical user interface)などの user interface と共に使われていることの多い

語を含むため、LSI では、上位になった。別の言い方をすると、user interface という用語に詳しい人であれば、この用例は適切と思えるが、初学者や user interface という用語に詳しくない人には、適切でない用例に見える。

英語教員に上記のような自動生成した用例リストについての意見を求めたところ、次のような意見が得られた。

１．潜在意味解析（LSI）で生成したリストは専門用語や、あまり使われない用語には適切な用例を含むが、初学者には適切ではない用例を多く含む。
２．入力した語の順序によらない、用例リストを生成できるのは、KWIC よりは便利である。
３．初学者には、長い文（語数の多い文）は用例として不適切である。

これらの意見から、初学者には BOW、中級レベルの英語学習者や一般的な用語には TF-IDF, 専門的な表現や用語には LSI が適していることが推測できた。

４．３　まとめと問題点

潜在意味解析を利用した用例の自動生成システムを作成し、実際に生成した用例リストについて、評価検討した。当初の計画では、大規模テキストデータから自動生成した用例リストを、学習者が実際に評価する予定だったが、次の二つの理由で、学習者による評価にはいたらなかった。

１．　潜在意味解析だけでは、初学者に適切な用例が生成できる可能性が少ないことが予想できた。
２．　頻度順などの潜在意味解析以外のアルゴリズムと組み合わせると、適切な用例が生成できる可能性が高いと推定できたが、推定を確認できるほどの十分な量のテキストデータの解析ができなかった。

２番目の理由の大きな原因は、使用したサーバーの処理能力の不足と、大規模な疎行列計算に関する経験と知識の不足だったと考える。サーバーの増強などの、研究方法の改善を行い、より信頼のおける評価と、幅広い層の学習者に対応できる用例の自動生成方法の実現を目指したい。

参考文献

[1]　http://corpus.byu.edu/coca/
[2]　I.S.P. Nation, Learning Vocabulary in Another Language, Cambridge University Press (2001)
[3]　I.S.P ネーション, “英語教師のためのボキャブラリーラーニング”, 松柏社, 2005.
[4]　Batia Laufer, Geke C. Ravenhorst-Kalovski “Lexical threshold revisited: Lexical text coverage, learners’ vocabulary size and reading comprehension”, Reading in a Foreign Language April 2010, Volume 22, No. 1
[5]大学英語教育学会基本語改訂委員会 , “大学英語教育学会基本語リスト JACET List of 8000 Basic Words”, 大学英語教育学会(2003).
[6] http://dumps.wikimedia.org/enwiki/20150403/
[7]　http://corpus.byu.edu/
[8] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor
[9]https://radimrehurek.com/gensim/index.html
[10]　Thomas K. Landauer et. al “Handbook of Latent Semantic Analysis” Psychology Press (2007)

５．主な発表論文等

〔学会発表〕（計　１　件）

来住 伸子, 岸 康人, 久島 智津子, 田近 裕子
「潜在意味解析などを利用した英文用例の自動生成」
第１４回情報科学技術フォーラム講演論文集、２０１５年９月　愛媛大学

〔その他〕
http://cooll.tsuda.ac.jp
この研究の先行研究として行った英語教材のソーシヤルブックマークサイト。

６．研究組織

(1)研究代表者
来住伸子　(KISHI, Nobuko)
津田塾大学・学芸学部・教授
研究者番号：50245990

(2)研究分担者
岸　康人　(KISHI, Yasuhito)
神奈川大学・ 総合理学研究所・研究員
研究者番号： 50552999

久島　智津子 (KUSHIMA, Chizuko)
津田塾大学・言語文化研究所・研究員
研究者番号： 80623876

田近　裕子　(TAJIKA, Hiroko)
津田塾大学・学芸学部・教授
研究者番号： 80188268